Image and Vision Computing 44 (2015) 15–28

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Word spotting in historical documents using primitive codebook and dynamic programming☆

Partha Pratim Roy *, Frédéric Rayar [1], Jean-Yves Ramel [1]

Laboratoire d'Informatique, Université François Rabelais, Tours, France

ARTICLE INFO

Article history:
Received 13 November 2013
Received in revised form 4 September 2015
Accepted 21 September 2015
Available online 22 October 2015

Keywords:
Word spotting
Document indexing
Approximate string matching
Coarse-to-fine

ABSTRACT

Word searching and indexing in historical document collections are a challenging problem because text characters are often touching or broken due to degradation or aging effects. In this paper, we present a novel approach towards word spotting using text line decomposition into character primitives and string matching. The text lines are initially separated by a segmentation process. Then each text line is described as sequences of primitive labels which correspond to single characters or parts of characters. These representative primitives are considered from a codebook of shapes generated from training pages taken from the collection. During indexation, the text lines are transcribed into strings of primitives in off-line stage and stored in files. For this purpose, an efficient indexation strategy using multi-label approach is used by a combination of two-level analysis of the primitives: coarse and fine levels. During retrieval, the query word image is encoded into strings of coarse and fine primitives chosen according to the codebook. Finally, a dynamic programming method based on approximate string matching is used to find similar primitive sequences in the text lines from the collection in runtime. We present the experimental evaluation on datasets of real life document images, gathered from historical books of different scripts. Experimental results show that the method is robust in searching text in noisy documents.

1. Introduction

Text searching in historical document is getting popular in Document Image Analysis (DIA) research community due to its complexity and the growing necessity for accessing the content of digitized books. In recent years, mass digitization of historical documents in libraries, museums are being performed and this digital information is made available to users through web-portals. By these portals, users are restricted to access only to view the pages that were digitized. Searching with content information (e.g. word) is available only if the corresponding pages are transcribed. In historical documents, due to degradation occurred by aging, strains, repetitive use, etc., the character recognition is not an easy task. Proper extraction of characters in such documents for recognition purpose is difficult. Incorrect segmentation of severely touching or broken characters is still one of the main causes for segmentation based recognition approaches [1]. Most of the word segmentation methods use space analysis between characters [2]. Sometimes due to non-uniform spacing between characters and words, it is difficult to segment words perfectly. Also, it is noticed that some pages of a historical book may contain text of different fonts. Thus, the recognition method needs to be robust to word segmentation problem and to tackle different fonts. We show two examples of document images from our collection in Fig. 1 that illustrate some of the issues described above. Automatic text transcription, performed by the available commercial OCR systems in these books is not satisfactory until now. Also, manual transcription of the archive is not feasible due to the large volume of data.

When processing such degraded documents, word spotting [3–7] techniques, an alternative to OCR, are useful to search the possible instances of specific/query words. These approaches do not require the recognition of every letter of the query word or the target words and thus are capable of similar word retrieval in the presence of small distortions. The features are generally computed from the whole word and thus the methods look for similar features in the target images. One of the bottle-neck of these word spotting methods is that most of them require a word segmentation step prior to the matching. If the words are not segmented properly, the features in target image do not match, thus these words cannot be retrieved. To overcome this problem, recently some segmentation free methods [8,9] have been proposed but their computation cost are too much high to be used in a real application for searching.

The goal of this work is to propose an efficient indexing scheme that will be able to search the text information in historical archives better and faster. To overcome OCR limitation, we propose to use Query By Example (QBE) principle in such a way that the user query image can be searched efficiently in a large volume of historical documents. The retrieval of text information will be fast and it will help the user to browse relevant information by overcoming problems that restrict OCR processing to historical books. Our proposed approach tries to overcome

☆ This paper has been recommended for acceptance by Seong-Whan Lee.

* Corresponding author. Tel.: +33 247 361 432.
E-mail address: partha.roy@univ-tours.fr (P.P. Roy).
[1] Tel.: +33 247 361 432.

http://dx.doi.org/10.1016/j.imavis.2015.09.006
0262-8856/

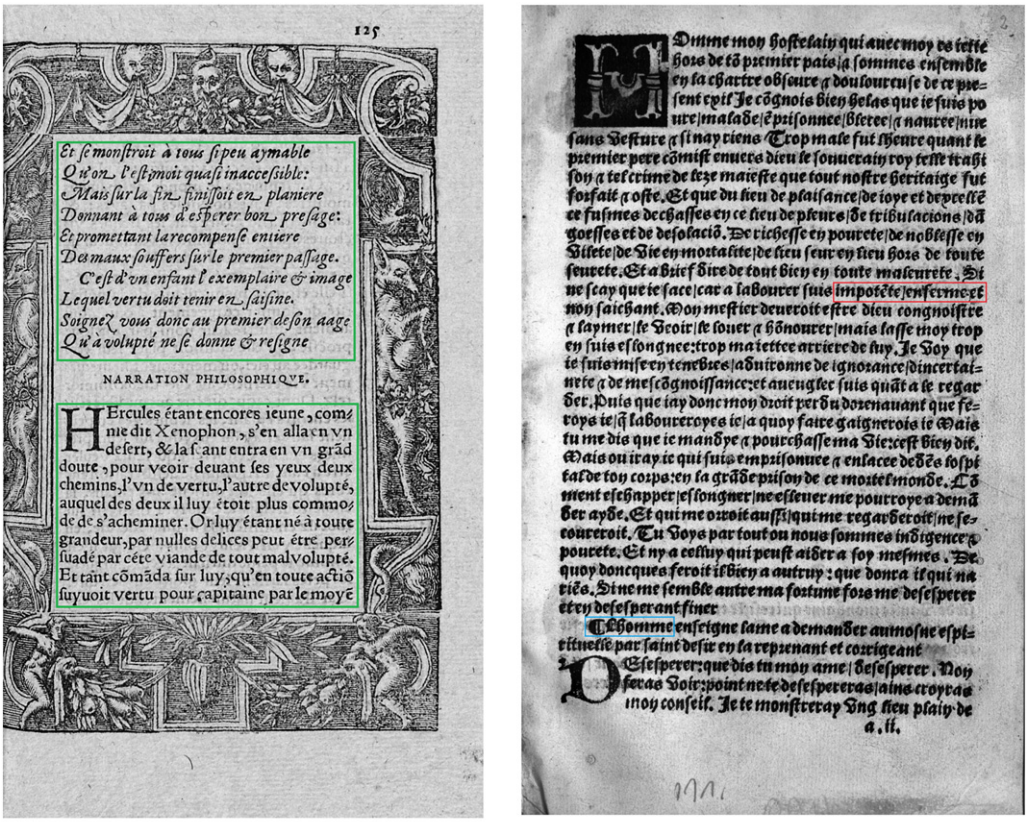

Fig. 1. Two documents from the historical collection (provided by CESR Tours) show some issues in recognition due to (i) difficulty to group characters in a word due of non-uniform spacing (Red box), (ii) character recognition problem because of strains and degradation (Blue box), (iii) different style of text information in the archive (Green box).

the difficulties of segmentation based word spotting methods by not requiring complete word segmentation before. Only, segmentation of text line that is relatively easier for layout segmentation of printed documents is considered in our approach. Also, the heavy searching-cost in segmentation-free word spotting methods is avoided by using a strategy of string encoding and matching of the text line image information.

Shape coding has been used efficiently to encode the words in printed documents [10]. Inspired with this idea, the proposed approach uses text primitive segmentation for word retrieval. With the same notion, we describe the text content (each text line) of historical books by basic feature shapes called primitive. A primitive consists of a single character or a part of a character. Primitive segmentation is performed using background information of the text image. To handle the background information, water reservoir concept [11] has been used. After the primitive extraction, similar primitives are grouped using a shape matching algorithm and a codebook of primitives is built. During indexation, the text contents in the book are encoded using the previously generated codebook of primitives. During the retrieval, a query word is also encoded by a string of primitives coming from the codebook. Next, a sub-string matching algorithm is applied to each encoded strings in the documents for retrieving query words. To make the retrieval process efficient, the encoding is done from two different codebooks of primitives : a coarse one corresponding to connected components and a fine one corresponding to glyphs (explained in Fig. 5). During the querying step, similarly, the coarse and the fine signatures are generated from the query image. The sub-string matching is performed by dynamic programming based approximate string matching algorithm. A bi-level matching is done to find similar words; using coarse approach first; and fine approach from the predetermined hypothetical locations only if necessary. This work is motivated by our preliminary work presented in [12,13]. The current work is an improved version with more details and an exhaustive experimentation has been performed. Numerous experiments have been performed to understand the different aspects of the methodology.

Our approach considers each text line as input for word spotting because, the segmentation of text lines in printed historical documents is relatively easier. Thus, we avoid the most difficult task of exact word segmentation in a document. The main contributions of this paper are the use of coarse and fine level text portions (primitives) instead of the whole word and encoding the text using these primitives for indexing. One advantage is that it searches for possible words in an efficient way using coarse level of primitive shapes (i.e. connected component) first. Then, if necessary, it uses fine primitives to detect strings of touching and broken characters. This two-level searching is robust to degradation such as touching or broken characters as we use fine level matching when coarse level matching fails. As the method searches for query word at the string level, using string matching in terms of primitives, response time is faster. The proposed approach can be applied in different scripts as the method uses dynamic codebook vocabulary for text encoding. Fig. 2 shows some examples of retrieval of

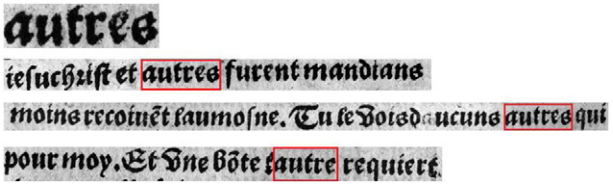

Fig. 2. Retrieval results of our system in a book where segmentation based word-spotting may not work.